Applied Mathematical Modelling 47 (2017) 276–293

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

A queueing system with decomposed service and inventoried preliminary services

Gabi Hanukov[a], Tal Avinadav[a,*], Tatyana Chernonog[a], Uriel Spiegel[a,b], Uri Yechiali[c]

[a] Department of Management, Bar-Ilan University, Ramat Gan 5290002, Israel
[b] Department of Economics, University of Pennsylvania, Philadelphia, USA
[c] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997801, Israel

ARTICLE INFO

Article history:
Received 5 September 2016
Revised 15 February 2017
Accepted 6 March 2017
Available online 12 March 2017

Subject classifications:
Queueing systems
Inventory control
Performance measures
Economic analysis
Matrix geometric
Laplace-Stieltjes transform

ABSTRACT

We study a single-server queue in which the service consists of two independent stages. The first stage is generic and can be performed even in the absence of customers, whereas the second requires the customer to be present. When the system is empty of customers, the server produces an inventory of first-stage ('preliminary') services (denoted PSs), which is used to reduce customers' overall sojourn times. We formulate and analyze the queueing-inventory system and derive its steady-state probabilities by using the matrix geometric method, which is based on calculating the so called rate matrix *R*. It is shown that the system's stability is not affected by the production rate of PSs, and that there are cases in which utilizing the server's idle time to produce PSs actually increases the fraction of time during which the server is dormant. A significant contribution is the derivation of an explicit expression of *R*, whose entries are written in terms of Catalan numbers. This type of result is rare in the literature and enables large-scale problems to be solved with low computational effort. Furthermore, by utilizing Laplace–Stieltjes transform and its inverse, we obtain the distribution function of customers' sojourn time. Finally, based on the probabilistic study, we carry out an economic analysis using a practical example from the fast food industry.

1. Introduction

In stochastic service systems, customer waiting time is commonly considered to be a reflection of system performance. Indeed, waiting time can directly affect the service provider's bottom line: Long waiting times may lead to a loss of customer goodwill, causing customers to abandon the service; moreover, the service provider may incur costs in an attempt to compensate customers for long delays. A common method of reducing customer waiting time is to increase the system's service capacity, either by using a more efficient server or by shifting to a multi-server system. However, adoption of either approach is likely to increase the service provider's operating costs. Accordingly, managers face the challenge of adapting their service processes in a way that reduces customers' waiting times, without substantially increasing their expenditures.

[*] Corresponding author.

E-mail addresses: german.khanukov@live.biu.ac.il (G. Hanukov), tal.avinadav@biu.ac.il (T. Avinadav), tatyana.chernonog@biu.ac.il (T. Chernonog), uriel.spiegel@biu.ac.il (U. Spiegel), uriy@post.tau.ac.il (U. Yechiali).

http://dx.doi.org/10.1016/j.apm.2017.03.008

In this paper, we present a novel approach to address this challenge, and we evaluate its performance using a single-server model. Our approach is relevant to service systems in which the service consists of two separate tasks: One task is generic, identical for all customers, and can be performed even when no customers are present; the other involves tailoring the service according to a particular customer's specific requirements, such that it can only be carried out when a customer is present in the system. We refer to the generic component of the service as a *preliminary service* (PS) and to the tailored component as a *complementary service* (CS). We propose that, in order to provide faster service to customers, the server can utilize the periods in which no customers are present in the system (referred to as the "server's idle time") to produce and store PSs for future arrivals. This way, upon arrival, a customer does not need to wait for the entire service to be carried out from the beginning: Rather, the server can withdraw a PS from inventory and subsequently perform a CS (which, presumably, takes less time than the full service), such that the customer leaves the system more quickly.

The proposed model is of broad interest and application. For instance, it can be used to improve assemble-to-order or co-production systems. An example of such a system is a bicycle shop, in which the first part of the service is assembling a bicycle's basic parts, and the second part is making specific adjustments in accordance with the customer's instructions and preferences (e.g., adding lights or a bell; adjusting the seat and pedals, etc.). Another domain of interest is the food service industry (including restaurants, fast food chains, and delis), in which, in many cases, basic elements of a dish are prepared in advance, and the dish is completed to order upon the customer's arrival. For example, in a pizzeria, servers can prepare and store plain (unbaked) pizza pies while the restaurant is empty, and when a customer arrives, they can add toppings per the customer's request and put the pizza in the oven. Considering that the global revenue of the fast food industry amounts to hundreds of billions of dollars per year (http://www.restaurant.org/News-Research/Research/Facts-at-a-Glance), even a small improvement in its efficiency has the potential to generate vast savings.

In what follows, after formulating our model of a two-stage service system with inventoried PSs, we explicitly derive performance measures for this system, including the distribution function of customers' sojourn time, mean queue length, and mean inventory level of PSs. The primary innovation of the model—inventoried PSs—introduces new questions that should be addressed. For example, is the quality of a CS identical to that of a service that is provided continuously from the start? What is the cost associated with storing PSs? How many PSs should be stored? Using a practical application of our model, we refer to those questions and provide managerial insights based on economic analysis and optimization.

Utilization of servers' idle time has long been discussed in the context of so-called 'vacation models', in which an idle server may be assigned to perform ancillary functions. This topic is thoroughly investigated in Levy and Yechiali [1,2], Kella and Yechiali [3], Takagi [4], Rosenberg and Yechiali [5], Boxma et al. [6], Yechiali [7], Lin and Ke [8], Jain and Jain [9], Ke et al. [10], Mytalas and Zazanis [11] and Guha et al. [12]. For situations in which operating costs are high, Yadin and Naor [13] propose an '*N*-policy vacation-type model', where an idle server should remain idle until the queue is of size *N*. Kella [14], Moreno [15], Lee and Yang [16], Lim et al. [17], Wei et al. [18], Yang and Wu [19], and Haridass and Arumuganathan [20] further developed this policy. Another collection of works that is closely related to our study investigates queues in which the service time of each individual customer is composed of multiple phases (stages). Choudhury and Madan [21], Choudhury [22], Choudhury et al. [23], Choudhury [24], and Choudhury and Deka [25] study M/G/1 and M^X/G/1 queues with two phases of heterogeneous service under different vacation policies.

The model we formulate is based on a quasi-birth-and-death (QBD) process, and we solve it using the matrix geometric method. The main contributions of this study, on top of the model formulation, are the following:

(i) Calculating explicitly all the entries of the rate matrix *R*, and showing the entries' relation to Catalan numbers.
(ii) Showing that the condition for system stability is independent of the PS production rate.
(iii) Deriving the probability density function (PDF) of a customer's sojourn time.
(iv) Showing how this model can be applied to a real-life problem, and providing means for economical optimization.

The remainder of this paper is organized as follows: In Section 2, we formulate the queueing-inventory model as a multi-dimensional Markovian process, and construct its infinitesimal generator matrix. In Section 3, we obtain the stability condition of the system, and calculate explicitly all the entries of the rate matrix *R*, from which all steady-state probabilities are derived. Section 4 presents various measures of system performance. In Section 5, we compute the fraction of time during which the server remains idle in our model, and compare it with the case in which no PSs are stored (i.e., the server performs a complete service from scratch for each arriving customer). In Section 6, we study the distribution of customer sojourn time. Section 7 provides analytical results of the system's performance measures for small values of PS inventory capacity. In Section 8, we provide managerial implications based on a practical example from the fast food industry. Finally, Section 9 provides concluding remarks and directions for future research.

2. Model formulation

We assume that customers arrive at a single-server queueing system according to a Poisson process with rate λ. An individual customer's service duration is composed of two independent stages. The first stage (stage 1, PS) can be carried out in the absence of customers. Thus, when no customers are present, the server produces and stores PSs for future use. When a customer arrives, if PSs are available in inventory, the server withdraws one PS and proceeds to complete the second stage of service (stage 2, CS). If no inventoried PSs are available when the customer arrives, the server performs stage 1 and stage 2 consecutively.

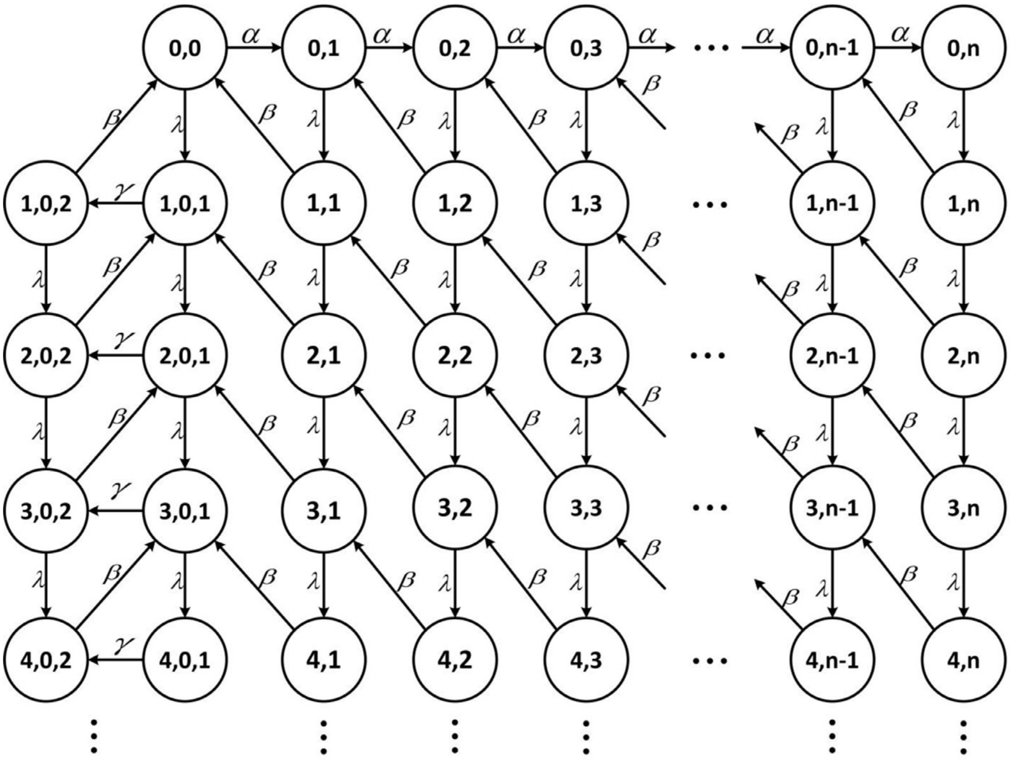

Fig. 1. Transition-rate diagram of the queueing-inventory system's states.

The duration of each stage of service is exponentially distributed. It should be noted that the duration of stage 1 when a customer is present might differ from its duration when a customer is absent: For example, when the customer is present, he might interfere with the server's work (e.g., by talking to the server) and slow it down, or he might make the server work more rapidly as a result of the supervision effect. Moreover, when the server performs stage 1 with the knowledge that it will be immediately followed by stage 2 (i.e., in the customer's presence), the server might adopt a different procedure than when producing a PS for inventory: In the example of the pizzeria, a server preparing plain pies for later use must take them to the refrigerator, whereas when preparing a pizza for immediate consumption (i.e., in the presence of a customer) the server proceeds directly to add toppings and put the pie in the oven. Thus, the mean duration of stage 1 is $1/\alpha$ when the customer is absent, and $1/\gamma$ when the customer is present. The mean duration of stage 2 is $1/\beta$. We assume that the maximum number of inventoried PSs is limited to n. When the inventory level reaches n and no customers are present, the server stops producing PSs and stays dormant. On the other hand, when the server is in the middle of producing a PS and a customer arrives, the server stops producing the PS, and immediately starts serving the newly-arrived customer.

We formulate the process as a quasi birth-and-death (QBD) process and analyze the queueing-inventory system in steady state. Let $L \in \{0,1,2,\ldots\}$ denote the number of customers present in the system, and let $S \in \{0,1,2,\ldots,n\}$ denote the number of PSs in the system. Let $H \in \{1,2\}$ denote the stage of service being provided to the customer at the front of the queue. Consequently, the state variable is three-dimensional when $S = 0$ and $L \geq 1$, and is two-dimensional otherwise. We define the following steady-state probabilities: $p_{i,0,k} = \Pr(L = i, S = 0, H = k)$, $i = 1, 2, \ldots, \infty$, $k = 1, 2$; $p_{0,0} = \Pr(L = 0, S = 0)$; $p_{i,j} = \Pr(L = i, S = j)$, $i = 0, 1, 2, \ldots, \infty$, $j = 1, 2, \ldots, n$. Fig. 1 depicts the transition rate diagram of the queueing-inventory system's states.

Two main methods are commonly used to obtain the system's steady-state probabilities: (i) probability-generating functions (PGFs; see, e.g., Litvak and Yechiali [26], Perel E. and Yechiali [27], Perel N. and Yechiali [28]), and (ii) matrix geometric analysis (see, e.g., Neuts [29], Latouche and Ramaswami [30]). The first method is based on constructing a set of $n + 1$ linear equations, whose unknowns are a set of PGFs that we seek to obtain, where each PGF corresponds to the state probabilities of a column in Fig. 1. By calculating the roots of a matrix related to the above finite set of linear equations, the so called 'boundary probabilities' are obtained and the PGFs are derived. In this paper, we use the second method, since, according to Hanukov et al. [31], it exhibits a computational advantage over the PGFs method.

To this end, we arrange the system's states in the following order:

$$\{(0,0),(0,1),\ldots,(0,n);(1,0,2),(1,0,1),(1,1),\ldots,(1,n);\ldots;(i,0,2),(i,0,1),(i,1),\ldots,(i,n);\ldots\},\quad i = 1,2,3,\ldots,$$

and construct its infinitesimal generator matrix (see, e.g., Neuts [29], p. 82, Ma et al. [32], and Zhou et al. [33]), denoted by Q, as

$$Q = \begin{pmatrix} B_0 & B_1 & 0 & 0 & 0 & \cdots \\ B_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1}$$

where the matrix B_0 (of order $(n+1) \times (n+1)$) is

$$B_0 = \begin{pmatrix} -(\lambda+\alpha) & \alpha & 0 & \cdots & 0 & 0 \\ 0 & -(\lambda+\alpha) & \alpha & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & & -(\lambda+\alpha) & \alpha \\ 0 & 0 & 0 & \cdots & 0 & -\lambda \end{pmatrix}, \tag{2}$$

the matrix B_1 (of order $(n+1) \times (n+2)$) is

$$B_1 = \begin{pmatrix} 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & & \lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}, \tag{3}$$

the matrix B_2 (of order $(n+2) \times (n+1)$) is

$$B_2 = \begin{pmatrix} \beta & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \beta & 0 & & 0 & 0 \\ 0 & \beta & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta & 0 \end{pmatrix}, \tag{4}$$

and the matrices A_0, A_1 and A_2 (each of order $(n+2) \times (n+2)$) are

$$A_0 = \begin{pmatrix} \lambda & 0 & \cdots & 0 & 0 \\ 0 & \lambda & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & & \lambda & 0 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}, \tag{5}$$

$$A_1 = \begin{pmatrix} -(\lambda+\beta) & 0 & 0 & \cdots & 0 & 0 \\ \gamma & -(\lambda+\gamma) & 0 & \cdots & 0 & 0 \\ 0 & 0 & -(\lambda+\beta) & & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & & -(\lambda+\beta) & 0 \\ 0 & 0 & 0 & \cdots & 0 & -(\lambda+\beta) \end{pmatrix}. \tag{6}$$

and

$$A_2 = \begin{pmatrix} 0 & \beta & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \beta & 0 & \cdots & 0 & 0 \\ 0 & 0 & \beta & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \beta & 0 \end{pmatrix}. \tag{7}$$

The matrices A_0, A_1 and A_2 will be used to derive the condition for system's stability, as described below.

3. Analysis

Let $\vec{e}_g \equiv (1, 1, ..., 1)^T$ denote a column vector of ones, where g is the dimension of $\vec{e}$, and let

$$A \equiv A_0 + A_1 + A_2 = \begin{pmatrix} -\beta & \beta & 0 & \cdots & 0 & 0 \\ \gamma & -\gamma & 0 & \cdots & 0 & 0 \\ 0 & \beta & -\beta & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \cdots & \beta & -\beta \end{pmatrix}. \tag{8}$$

Theorem 1. *The queuing-inventory system is stable if and only if* $1/\lambda > 1/\gamma + 1/\beta$.

Proof. According to Neuts [26, p. 83], the stability condition is

$$\vec{\pi}\, A_0\, \vec{e}_{n+2} < \vec{\pi}\, A_2\, \vec{e}_{n+2}, \tag{9}$$

where $\vec{\pi} \equiv (\pi_{01}, \pi_{02}, \pi_1, ..., \pi_n)$ is the unique solution of the linear system

$$\begin{aligned} &\vec{\pi}\, A = \vec{0}, \\ &\vec{\pi}\, \vec{e}_{n+2} = 1. \end{aligned} \tag{10}$$

From Eq. (10), we get $\vec{\pi} = (\gamma/(\gamma+\beta), \beta/(\gamma+\beta), 0, ..., 0)$, so that Eq. (9) translates into $1/\lambda > 1/\gamma + 1/\beta$. □

Theorem 1 has the following intuitive explanation: the server produces CSs at rate β only when PSs are available. When no PSs are available, the system shifts to an M/G/1-type queue with service duration composed of the sum of two independent and not necessarily identical exponential random variables with overall mean service time $1/\gamma + 1/\beta$. Thus, for a sufficiently large number of customers in the system, the entire inventory of PSs will be exhausted, and the system will behave as an M/G/1 queue. The corresponding stability condition is $\lambda(1/\gamma + 1/\beta) < 1$, independent of the PS production rate, α.

The so called rate matrix $R = [r_{i,j}]$ (of order $(n+2) \times (n+2)$) satisfies

$$A_0 + RA_1 + R^2 A_2 = 0_{n+2,n+2}. \tag{11}$$

In Appendix A, we give explicitly all the entries of the left-hand side of Eq. (11). The solution of Eq. (11) may result in several values for each entry in R, but only the minimal non-negative value is taken (Neuts [29], p. 82). Usually, $r_{i,j}$ values are obtained numerically (Latouche and Ramaswami [30], chapter 8) by *successive substitutions* (Neuts [29], p. 37). In contrast, as shown in the following theorem, we have successfully obtained closed-form expressions for *all* $r_{i,j}$. Such an explicit *complete* solution is rare.

Theorem 2.

$$r_{i,1} = \begin{cases} \lambda/\beta & 1 \le i \le 2 \\ \frac{C_{i-2}\beta^{i-4}\lambda^{i-1}}{(\lambda+\beta)^{2i-5}} + \sum\limits_{k=3}^{i-1} \frac{C_{i-k}\beta^{i-k-1}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)}} r_{k,1} & 3 \le i \le n+2 \end{cases}, \tag{12}$$

$$r_{i,2} = \begin{cases} \lambda^2/(\gamma\beta) & i = 1 \\ \lambda(\lambda+\beta)/(\gamma\beta) & i = 2 \\ \frac{C_{i-2}\beta^{i-4}\lambda^{i-1}}{\gamma(\lambda+\beta)^{2i-6}} + \sum\limits_{k=3}^{i-1} \frac{C_{i-k}\beta^{i-k-1}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)}} r_{k,2} & 3 \le i \le n+2 \end{cases}, \tag{13}$$

$$r_{i,j} = \begin{cases} 0 & 3 \le j \le n+2,\ i < j \\ \frac{C_{i-j}\beta^{i-j}\lambda^{i-j+1}}{(\lambda+\beta)^{2(i-j)+1}} & 3 \le j \le i \le n+2 \end{cases}, \tag{14}$$

where

$$C_m = \frac{(2m)!}{(m+1)!\,m!} \tag{15}$$

is the mth Catalan number (Koshy [34]), $m = 0, 1, 2, ...,$

Proof. See Appendix B.

For example, when $n = 6$,

$$R = \begin{pmatrix} \frac{\lambda}{\beta} & \frac{\lambda^2}{\gamma\beta} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{\lambda}{\beta} & \frac{\lambda(\beta+\lambda)}{\gamma\beta} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{\lambda^2}{\beta(\lambda+\beta)} & \frac{\lambda^2}{\gamma\beta} & \frac{\lambda}{\lambda+\beta} & 0 & 0 & 0 & 0 & 0 \\ \frac{\lambda^3(2\beta+\lambda)}{\beta(\lambda+\beta)^3} & \frac{\lambda^3(2\beta+\lambda)}{\gamma\beta(\lambda+\beta)^2} & \frac{\beta\lambda^2}{(\lambda+\beta)^3} & \frac{\lambda}{\lambda+\beta} & 0 & 0 & 0 & 0 \\ \frac{\lambda^4(5\beta^2+4\beta\lambda+\lambda^2)}{\beta(\lambda+\beta)^5} & \frac{\lambda^4(5\beta^2+4\beta\lambda+\lambda^2)}{\gamma\beta(\lambda+\beta)^4} & \frac{2\beta^2\lambda^3}{(\lambda+\beta)^5} & \frac{\beta\lambda^2}{(\lambda+\beta)^3} & \frac{\lambda}{\lambda+\beta} & 0 & 0 & 0 \\ \frac{\lambda^5(14\beta^3+14\beta^2\lambda+6\beta\lambda^2+\lambda^3)}{\beta(\lambda+\beta)^7} & \frac{\lambda^5(14\beta^3+14\beta^2\lambda+6\beta\lambda^2+\lambda^3)}{\gamma\beta(\lambda+\beta)^6} & \frac{5\beta^3\lambda^4}{(\lambda+\beta)^7} & \frac{2\beta^2\lambda^3}{(\lambda+\beta)^5} & \frac{\beta\lambda^2}{(\lambda+\beta)^3} & \frac{\lambda}{\lambda+\beta} & 0 & 0 \\ \frac{\lambda^6(42\beta^4+48\beta^3\lambda+27\beta^2\lambda^2+8\beta\lambda^3+\lambda^4)}{\beta(\lambda+\beta)^9} & \frac{\lambda^6(42\beta^4+48\beta^3\lambda+27\beta^2\lambda^2+8\beta\lambda^3+\lambda^4)}{\gamma\beta(\lambda+\beta)^8} & \frac{14\beta^4\lambda^5}{(\lambda+\beta)^9} & \frac{5\beta^3\lambda^4}{(\lambda+\beta)^7} & \frac{2\beta^2\lambda^3}{(\lambda+\beta)^5} & \frac{\beta\lambda^2}{(\lambda+\beta)^3} & \frac{\lambda}{\lambda+\beta} & 0 \\ \frac{\lambda^7\begin{pmatrix}132\beta^5+165\beta^4\lambda+110\beta^3\lambda^2\\+44\beta^2\lambda^3+10\beta\lambda^4+\lambda^5\end{pmatrix}}{\beta(\lambda+\beta)^{11}} & \frac{\lambda^7\begin{pmatrix}132\beta^5+165\beta^4\lambda+110\beta^3\lambda^2\\+44\beta^2\lambda^3+10\beta\lambda^4+\lambda^5\end{pmatrix}}{\gamma\beta(\lambda+\beta)^{10}} & \frac{42\beta^5\lambda^6}{(\lambda+\beta)^{11}} & \frac{14\beta^4\lambda^5}{(\lambda+\beta)^9} & \frac{5\beta^3\lambda^4}{(\lambda+\beta)^7} & \frac{2\beta^2\lambda^3}{(\lambda+\beta)^5} & \frac{\beta\lambda^2}{(\lambda+\beta)^3} & \frac{\lambda}{\lambda+\beta} \end{pmatrix}.$$

Let $\vec{p}_0 \equiv (p_{0,0}, p_{0,1}, \ldots, p_{0,n-1}, p_{0,n})$ and $\vec{p}_i \equiv (p_{i,0,2}, p_{i,0,1}, p_{i,1}, p_{i,2}, \ldots, p_{i,n-1}, p_{i,n})$, $i = 1, 2, \ldots, \infty$. In order to calculate $p_{i,\,0,\,k}$ and $p_{i,\,j}$, $i = 0, 1, \ldots, \infty$, $j = 0, 1, \ldots, n$, $k = 1, 2$, the vector of boundary probabilities $\vec{p}_0$ and the vector $\vec{p}_1$ have to be obtained. This is accomplished as follows: Set $\vec{p} \equiv (\vec{p}_0, \vec{p}_1, \vec{p}_2, \ldots)$. Then the probability vectors uniquely satisfy $\vec{p}Q = 0$ and $\vec{p}_0\vec{e}_{n+1} + \sum_{i=1}^{\infty} \vec{p}_i\vec{e}_{n+2} = 1$. It follows that $\vec{p}_0$ and $\vec{p}_1$ are calculated by:

$$\begin{aligned} &\vec{p}_0 B_0 + \vec{p}_1 B_2 = \vec{0}_{n+1} \\ &\vec{p}_0 B_1 + \vec{p}_1 [A_1 + RA_2] = \vec{0}_{n+2} \\ &\vec{p}_0 \vec{e}_{n+1} + \vec{p}_1 [I - R]^{-1} \vec{e}_{n+2} = 1. \end{aligned} \tag{16}$$

Then, all other steady-state probabilities are obtained by

$$\vec{p}_i = \vec{p}_1 R^{i-1}, \quad i = 1, 2, 3, \ldots, \infty. \tag{17}$$

4. Performance measures

For a given capacity of n PSs, let $L(n)$ and $L_q(n)$ denote the mean number of customers in the system and in queue, respectively. Similarly, let $W(n)$ and $W_q(n)$ denote, respectively, the mean sojourn time of a customer in the system and in queue; $S(n)$ and $S_q(n)$ denote, respectively, the mean number of PSs in the system and in inventory. Finally, let $T(n)$ and $T_q(n)$ denote, respectively, the mean time a PS resides in the system and in inventory.

Using Eq. (17), we have:

$$L(n) = \sum_{i=1}^{\infty} i\left(\vec{p}_i \vec{e}_{n+2}\right) = \sum_{i=1}^{\infty} i\left(\vec{p}_1 R^{i-1} \vec{e}_{n+2}\right) = \vec{p}_1 \left(\sum_{i=1}^{\infty} iR^{i-1}\right) \vec{e}_{n+2}.$$

Since $\sum_{i=1}^{\infty} iR^{i-1} = ([I-R]^{-1})^2 = [I-R]^{-2}$, then

$$L(n) = \vec{p}_1 [I-R]^{-2} \vec{e}_{n+2}. \tag{18}$$

Since $\vec{p}_0 \vec{e}_{n+1}$ is the probability that the system is empty of customers, we readily obtain

$$L_q(n) = L(n) - \left(1 - \vec{p}_0 \vec{e}_{n+1}\right). \tag{19}$$

By Little's law,

$$W(n) = L(n)/\lambda, \tag{20}$$

$$W_q(n) = L_q(n)/\lambda. \tag{21}$$

Fig. 2 depicts the mean customer sojourn time as a function of the PS capacity n for $\lambda = 28$, $\gamma = 18$, $\beta = 22.5$ and $\alpha = \{18, 30\}$. It is observed that $W(n)$ is higher for the lower value of α, and that it is a monotone decreasing convex function of n with an asymptotic value. For example, for $\alpha = 30$, $W(0) = 0.401 > W(5) = 0.166 > W(10) = 0.094$, and so on. When n grows, the impact diminishes to an asymptotic value of approximately 0.069, which further justifies the policy of bounding the value of n.

To calculate $S(n)$, we define two vectors $\vec{v} \equiv (0, 1, 2, \ldots, n)^T$ and $\vec{w} \equiv (0, 0, 1, 2, \ldots, n)^T$, so that

$$S(n) = \sum_{i=0}^{\infty} \sum_{j=1}^{n} j p_{i,j} = \vec{p}_0 \vec{v} + \sum_{i=1}^{\infty} \vec{p}_i \vec{w} = \vec{p}_0 \vec{v} + \vec{p}_1 \left(\sum_{i=1}^{\infty} R^{i-1}\right) \vec{w} = \vec{p}_0 \vec{v} + \vec{p}_1 [I-R]^{-1} \vec{w}. \tag{22}$$

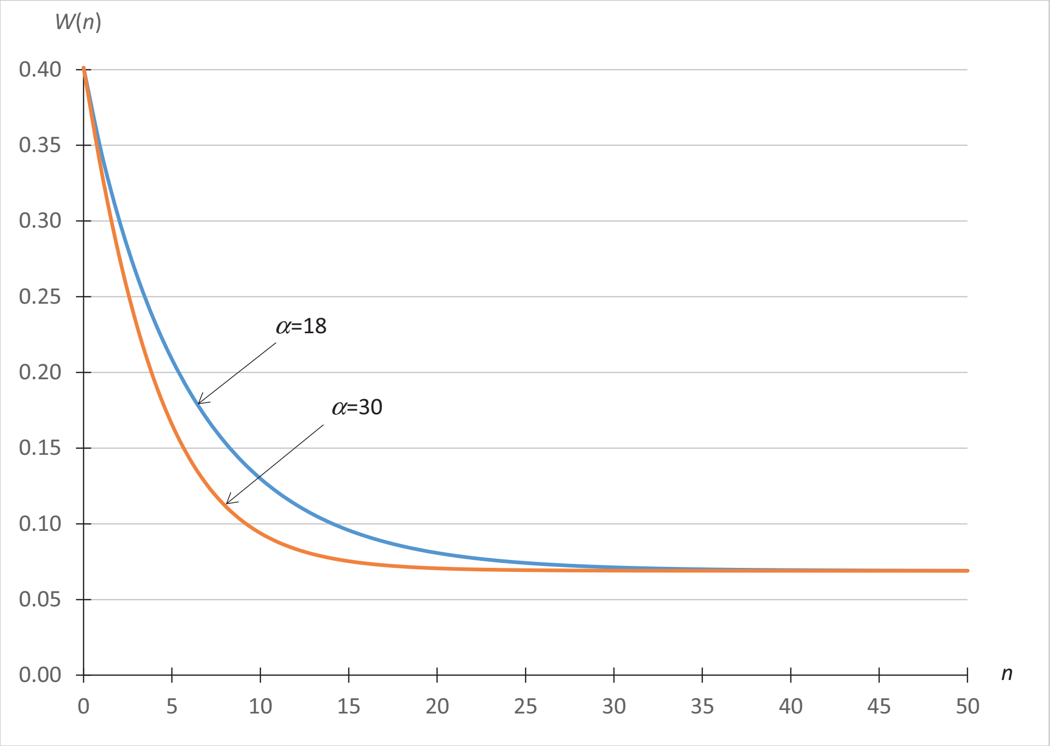

Fig. 2. Plots of $W(n)$ for $\lambda = 28$, $\gamma = 18$, $\beta = 22.5$ and $\alpha = \{18,\ 30\}$.

Similarly, to calculate $S_q(n)$, we define a vector $\vec{u} \equiv (0, 0, 0, 1, 2, ..., n-1)^T$ for $n \geq 1$, so that

$$S_q(n) = \sum_{j=1}^{n} j p_{0,j} + \sum_{i=1}^{\infty}\sum_{j=1}^{n}(j-1)p_{i,j} = \vec{p}_0\vec{v} + \sum_{i=1}^{\infty}\vec{p}_i\vec{u} = \vec{p}_0\vec{v} + \vec{p}_1\left(\sum_{i=1}^{\infty}R^{i-1}\right)\vec{u} = \vec{p}_0\vec{v} + \vec{p}_1[I-R]^{-1}\vec{u}. \tag{23}$$

The effective production rate of PSs (produced only when the system is empty and there are fewer than n inventoried PSs) is

$$\alpha_{eff}(n) = \alpha\left(\vec{p}_0\,\vec{e}_{n+1} - p_{0,n}\right). \tag{24}$$

By Little's law,

$$T(n) = S(n)/\alpha_{eff}(n), \tag{25}$$

$$T_q(n) = S_q(n)/\alpha_{eff}(n). \tag{26}$$

5. The effect of service decomposition on the server's idleness

For each row in Fig. 1, define its marginal probability:$p_{0,\bullet} = \sum_{j=0}^{n} p_{0,j}$; $p_{i,\bullet} = \sum_{j=1}^{n} p_{i,j}$, $i = 1, 2, ..., \infty$. Similarly, for each column, define its marginal probability:$p_{\bullet,0,k} = \sum_{i=1}^{\infty} p_{i,0,k}$, $k = 1, 2$; and $p_{\bullet,j} = \sum_{i=0}^{\infty} p_{i,j}$, $j = 1...n$. Then, for $i = 0$, writing the balance equation for each state $(0, j)$, $j = 0,1,2,...,n$, and then summing over j, leads to a balance equation of the horizontal cut between row 0 and row 1:

$$\lambda p_{0,\bullet} = \beta(p_{1,0,2} + p_{1,\bullet}). \tag{27}$$

Similarly, for each row $i = 1,2,3,..,\infty$, writing the balance equation for each state and summing along the row leads to a balance equation of the horizontal cut between row i and row $i+1$:

$$\lambda(p_{i,0,1} + p_{i,0,2} + p_{i,\bullet}) = \beta(p_{i+1,0,2} + p_{i+1,\bullet}). \tag{28}$$

Consider now the columns of Fig. 1. Writing the balance equation for each state along the left column, and summing them (which amounts to a vertical cut between the first two columns) leads to:

$$\gamma p_{\bullet,0,1} = \beta p_{\bullet,0,2}. \tag{29}$$

Continuing with cuts between all other columns, j and $j+1$, for $j = 0,1,2,\ldots,n-1$, we obtain the following balance equations:

$$\alpha p_{0,j} = \beta(p_{\bullet,j+1} - p_{0,j+1}). \tag{30}$$

Summing Eqs. (27) and (28) over $i = 1, 2, ..., \infty$ results in

$$\lambda\left(p_{0,\bullet} + \sum_{i=1}^{\infty}(p_{i,0,1} + p_{i,0,2} + p_{i,\bullet})\right) = \beta\left(p_{\bullet,0,2} + \sum_{i=1}^{\infty} p_{i,\bullet}\right). \tag{31}$$

Since $p_{0,\bullet} + \sum_{i=1}^{\infty}(p_{i,0,1} + p_{i,0,2} + p_{i,\bullet}) = 1$ and $p_{\bullet,0,2} + \sum_{i=1}^{\infty} p_{i,\bullet} = 1 - p_{\bullet,0,1} - p_{0,\bullet}$, then, using algebraic manipulations, Eq. (31) can be written as

$$p_{\bullet,0,1} = 1 - p_{0,\bullet} - \lambda/\beta. \tag{32}$$

Summing Eqs. (29) and (30) over $j = 0, 1, 2, ..., n-1$ yields

$$\gamma p_{\bullet,0,1} + \alpha(p_{0,\bullet} - p_{0,n}) = \beta\left(p_{\bullet,0,2} + \sum_{j=1}^{n} p_{\bullet,j} - \sum_{j=1}^{n} p_{0,j}\right). \tag{33}$$

Since $\sum_{j=1}^{n} p_{\bullet,j} = 1 - p_{0,0} - p_{\bullet,0,1} - p_{\bullet,0,2}$ and $\sum_{j=1}^{n} p_{0,j} = p_{0,\bullet} - p_{0,0}$, then, using algebraic manipulations, Eq. (33) can be rewritten as

$$\gamma p_{\bullet,0,1} + \alpha(p_{0,\bullet} - p_{0,n}) = \beta(1 - p_{\bullet,0,1} - p_{0,\bullet}). \tag{34}$$

Substituting Eq. (32) in (34) yields

$$\gamma(1 - p_{0,\bullet} - \lambda/\beta) + \alpha(p_{0,\bullet} - p_{0,n}) = \lambda, \tag{35}$$

which contains only the boundary probabilities $p_{0,\,j}$, $j = 0, 1, 2, ..., n$.

The following lemma gives the fraction of time in which the server is idle in an ordinary system that does not store PSs (i.e., $n = 0$).

Lemma 1. *When $n = 0$, the fraction of time in which the server is idle is $1 - \lambda(1/\gamma + 1/\beta)$.*

Proof. When $n = 0$ the system can be regarded as an M/G/1 queue with mean service time $1/\gamma + 1/\beta$. Indeed, from Fig. 1, the horizontal cuts yield $\lambda p_{0,\,0} = \beta p_{1,\,0,\,2}$ and $\lambda(p_{i,\,0,\,1} + p_{i,\,0,\,2}) = \beta p_{i\,+\,1,\,0,\,2}$, $i = 1, 2, ..., \infty$, while a vertical cut yields $\gamma p_{\bullet,\,0,\,1} = \beta p_{\bullet,\,0,\,2}$. Summing all the equations of the horizontal cuts and using the relation $p_{0,\,0} + p_{\bullet,\,0,\,1} + p_{\bullet,\,0,\,2} = 1$ results in $\lambda = \beta p_{\bullet,\,0,\,2}$. Thus, $p_{\bullet,\,0,\,1} = \lambda/\gamma$ and $p_{0,\,0} = 1 - \lambda(1/\gamma + 1/\beta)$. □

The following proposition compares the fraction of idle time in our queueing-inventory model with that obtained in a model in which there is no possibility to store PSs in inventory. When $n \geq 1$, the server's fraction of idle time is $p_{0,\,n}$, whereas when $n = 0$ it is given by Lemma 1.

Proposition 1. $\text{sgn}(p_{0,n} - (1 - \lambda(1/\gamma + 1/\beta))) = \text{sgn}(1/\gamma - 1/\alpha)$.

Proof. Eq. (35) can be written as $(\alpha - \gamma)p_{0,\,\bullet} = \alpha p_{0,\,n} + \lambda + \gamma(\lambda/\beta - 1)$. Dividing the latter equation by $\alpha\gamma$ yields

$$(1/\gamma - 1/\alpha)p_{0,\bullet} = (1/\gamma)p_{0,n} - (1/\alpha)(1 - \lambda(1/\gamma + 1/\beta)),$$

which can be written as

$$(1/\gamma - 1/\alpha)\sum_{j=0}^{n-1} p_{0,j} = (1/\alpha)(p_{0,n} - (1 - \lambda(1/\gamma + 1/\beta))).$$

The claim is proved since both $\alpha > 0$ and $\sum_{j=0}^{n-1} p_{0,j} > 0$.

Thus, the fraction of time during which the server is idle is not less than that in a system in which PSs cannot be stored *if and only if* the mean duration of the first stage of service when a customer is present is not less than that when the customer is absent. It is possible that $\alpha > \gamma$, due to a negative effect of the customer's presence on the server's production rate, as discussed in Section 2. This leads to a paradox: utilizing the server's idle time may increase the fraction of time in which the server is idle.

6. Laplace–Stieltjes transform (LST) of customers' sojourn time

Let $\tilde{\Lambda}(s) = \gamma/(\gamma + s)$ and $\tilde{B}(s) = \beta/(\beta + s)$ denote the LSTs of the first and second service stages, respectively. Let $\tilde{W}(s)$ denote the LST of a customer's sojourn time in the system with capacity n. Then, considering all state probabilities of the system, and the corresponding service stages that an arriving customer follows until leaving the system (see Fig. 1), we obtain

$$\tilde{W}(s) \equiv E[e^{-sW}] = p_{0,0}\tilde{\Lambda}(s)\tilde{B}(s) + \sum_{i=1}^{\infty} p_{i,0,1}\tilde{\Lambda}^{i+1}(s)\tilde{B}^{i+1}(s) \tag{36}$$

$$+ \sum_{i=1}^{\infty} p_{i,0,2}\tilde{\Lambda}^{i}(s)\tilde{B}^{i+1}(s) + \sum_{j=1}^{n}\left(\sum_{i=0}^{j-1} p_{i,j}\tilde{B}^{i+1}(s) + \sum_{i=j}^{\infty} p_{i,j}\tilde{\Lambda}^{i+1-j}(s)\tilde{B}^{i+1}(s)\right). \tag{36}$$

For example, if a customer finds the system in state (i, j), where $i \le j$, his or her total sojourn time in the system follows an Erlang distribution with $(i + 1)$ stages and rate parameter β. The balance equation $\vec{p}Q = 0$ can be written explicitly as follows:

For $j = 0$:

$$(\lambda + \alpha)p_{0,0} = \beta(p_{1,0,2} + p_{1,1}),\ \text{(for } i = 0), \tag{37}$$

$$(\lambda + \gamma)p_{1,0,1} = \lambda p_{0,0} + \beta(p_{2,0,2} + p_{2,1}),\ \text{(for } i = 1), \tag{38}$$

$$(\lambda + \beta)p_{1,0,2} = \gamma p_{1,0,1},\ \text{(for } i = 1), \tag{39}$$

$$(\lambda + \gamma)p_{i,0,1} = \lambda p_{i-1,0,1} + \beta(p_{i+1,0,2} + p_{i+1,1}),\ i \ge 2, \tag{40}$$

$$(\lambda + \beta)p_{i,0,2} = \lambda p_{i-1,0,2} + \gamma p_{i,0,1},\ i \ge 2. \tag{41}$$

For $1 \le j \le n - 1$:

$$(\lambda + \alpha)p_{0,j} = \alpha p_{0,j-1} + \beta p_{1,j+1},\ \text{(for } i = 0), \tag{42}$$

$$(\lambda + \beta)p_{i,j} = \lambda p_{i-1,j} + \beta p_{i+1,j+1},\ i \ge 1. \tag{43}$$

Finally, for $j = n$:

$$\lambda p_{0,n} = \alpha p_{0,n-1},\ \text{(for } i = 0), \tag{44}$$

$$(\lambda + \beta)p_{i,n} = \lambda p_{i-1,n},\ i \ge 1. \tag{45}$$

Remark. One can use Eqs. (37)–(45) to derive the finite set of linear equations for computing the PGFs (the alternative method of calculating steady-state probabilities discussed in Section 2).

Define: $D_{0,1}(s) = \sum_{i=1}^{\infty} p_{i,0,1}\tilde{\Lambda}^{i+1}(s)\tilde{B}^{i+1}(s)$; $D_{0,2}(s) = \sum_{i=1}^{\infty} p_{i,0,2}\tilde{\Lambda}^{i}(s)\tilde{B}^{i+1}(s)$, and $D_j(s) = \sum_{i=0}^{\infty} p_{i,j}\tilde{\Lambda}^{i}(s)\tilde{B}^{i}(s)$.

Then, by multiplying each equation in the set (37)–(45) by $\tilde{\Lambda}^{i}(s)\tilde{B}^{i}(s)$, summing the equations over all values of i for each value of j separately, and using some algebra, we obtain

$$\begin{aligned} j = 0, k = 1 :\quad & ((\lambda + \gamma)\tilde{B}(s) - \lambda\tilde{\Lambda}(s)\tilde{B}^2(s))D_{0,1}(s) - \beta D_{0,2}(s) - \beta\tilde{B}(s)D_1(s) \\ & = (\lambda(\tilde{\Lambda}(s)\tilde{B}(s) - 1) - \alpha)\tilde{\Lambda}(s)\tilde{B}^2(s)p_{0,0} - \beta\tilde{B}(s)p_{0,1}, \end{aligned} \tag{46}$$

$$j = 0, k = 2 :\quad \big((\lambda + \beta)\tilde{\Lambda}(s) - \lambda\tilde{\Lambda}^2(s)\tilde{B}(s)\big)D_{0,2}(s) - \gamma D_{0,1}(s) = 0, \tag{47}$$

$$1 \le j \le n-1: \qquad (\lambda(1-\tilde{\Lambda}(s)\tilde{B}(s))+\beta)\tilde{\Lambda}(s)\tilde{B}(s)D_j(s)-\beta D_{j+1}(s)$$
$$= \tilde{\Lambda}(s)\tilde{B}(s)\alpha p_{0,j-1}-\tilde{\Lambda}(s)\tilde{B}(s)(\alpha-\beta)p_{0,j}-\beta p_{0,j+1}, \tag{48}$$

$$j=n: \qquad \big(\lambda(1-\tilde{\Lambda}(s)\tilde{B}(s)\big)+\beta)D_n(s)=\alpha p_{0,n-1}+\beta p_{0,n}. \tag{49}$$

Eqs. (46)–(49) comprise a set of $n+2$ linear equations with unknowns $D_{0,\,1}(s)$, $D_{0,\,2}(s)$ and $D_j(s)$ for $j = 1, 2, ..., n$. To simplify the presentation of Eq. (36), we use the following relation:

$$\sum_{i=j}^{\infty} p_{i,j}\tilde{\Lambda}^{i+1-j}(s)\tilde{B}^{i+1}(s)=\tilde{\Lambda}^{1-j}(s)\tilde{B}(s)\sum_{i=j}^{\infty} p_{i,j}\tilde{\Lambda}^{i}(s)\tilde{B}^{i}(s)$$
$$=\tilde{\Lambda}^{1-j}(s)\tilde{B}(s)\left(\sum_{i=0}^{\infty} p_{i,j}\tilde{\Lambda}^{i}(s)\tilde{B}^{i}(s)-\sum_{i=0}^{j-1} p_{i,j}\tilde{\Lambda}^{i}(s)\tilde{B}^{i}(s)\right)=\tilde{\Lambda}^{1-j}(s)\tilde{B}(s)D_j(s)-\sum_{i=0}^{j-1} p_{i,j}\tilde{\Lambda}^{i+1-j}(s)\tilde{B}^{i+1}(s), \tag{50}$$

so

$$\tilde{W}(s)=p_{0,0}\tilde{\Lambda}(s)\tilde{B}(s)+D_{0,1}(s)+D_{0,2}(s)+\sum_{j=1}^{n}\left(\tilde{\Lambda}^{1-j}(s)\tilde{B}(s)D_j(s)+\sum_{i=0}^{j-1} p_{i,j}\tilde{B}^{i+1}(s)\big(1-\tilde{\Lambda}^{i+1-j}(s)\big)\right). \tag{51}$$

The explicit expression of $\tilde{W}(s)$ depends on the capacity n and is obtained by: (i) solving the set of Eqs. (46)–(49) and extracting $D_{0,\,1}(s)$, $D_{0,\,2}(s)$ and $D_j(s)$ for $j = 1, 2, ..., n$; (ii) calculating the probabilities $p_{i,\,j}$, $i = 0, 1, ..., n-1$, $j = i+1, i+2, ..., n$ given by Eqs. (16) and (17); and (iii) substituting the results of (i) and (ii) in Eq. (51).

7. Analytical results for small values of *n*

In what follows, we obtain explicit results for the system's performance measures, as well as for the PDF of a customer's sojourn time, for small values of n. Maple 2016 software enabled us to obtain closed-form expressions of the system's steady-state probabilities and performance measures up to $n = 18$. For larger values of n (up to 200), we were able to obtain numerically those probabilities and measures. Herein, we present results for $n = 0$, 1, and 2; for $n = 3, 4,..., 18$, the expressions are cumbersome and are thus not shown.

7.1. The case of n = 0

This case presents a system in which PSs cannot be stored in inventory, which results in an M/PH/1 queue, in which the service time consists of two stages, each of which has an exponentially-distributed duration (not necessarily with identical rate parameter). By Eqs. (18) and (19), we obtain the means:

$$L(0)=\frac{\lambda(\gamma+\beta-\lambda)}{\beta\gamma-\lambda(\gamma+\beta)}, \tag{52}$$

and

$$L_q(0)=\frac{\lambda^2(\gamma^2+\beta^2+\beta\gamma)}{\beta\gamma(\beta\gamma-\lambda(\gamma+\beta))}. \tag{53}$$

By Little's law, $W(1)$ and $W_q(1)$ are calculated by using Eqs. (20) and (21). By Eq. (51), the LST of the sojourn time is

$$\tilde{W}(s)=\frac{\beta\gamma-\lambda(\gamma+\beta)}{(\gamma+s)(\beta+s)-\lambda(\gamma+\beta+s)}. \tag{54}$$

For simplicity of presentation, let $\Psi \equiv \sqrt{(\beta-\gamma)^2+\lambda(\lambda+2(\gamma+\beta))}$. Using inverse Laplace transform, we explicitly obtain the PDF of a customer's sojourn time, W:

$$f_W(t)=L^{-1}\big(\tilde{W}(s)\big)=\frac{2(\beta\gamma-\lambda(\gamma+\beta))}{\Psi}e^{-0.5(\gamma+\beta-\lambda)\cdot t}\sinh(0.5\Psi\cdot t), \tag{55}$$

where $\sinh(x) = 0.5(e^x - e^{-x})$. In Fig. 3, we plot $f_W(t)$ for $n = 0$, $\lambda = 8$, $\alpha = 20$, $\gamma = 18$ and $\beta = 22.5$. It is observed that $f_W(t)$ is unimodal with a right tail decaying exponentially.

Notice that the case $n = 0$ is equivalent to an M/G/1 queue with service time composed of two independent exponential durations with parameters γ and β, respectively. The LST of the total service time is then given by $\tilde{B}(s)=\frac{\gamma}{\gamma+s}\frac{\beta}{\beta+s}$. Substituting $\tilde{B}(s)$ in the known formula (see, e.g., Cooper [35], p. 217) of the LST of the sojourn time in a regular M/G/1 queue, namely, $\tilde{W}(s)=(1-\lambda(\frac{1}{\gamma}+\frac{1}{\beta}))\frac{s\cdot\tilde{B}(s)}{\lambda\cdot\tilde{B}(s)+s-\lambda}$, results in Eq. (54).

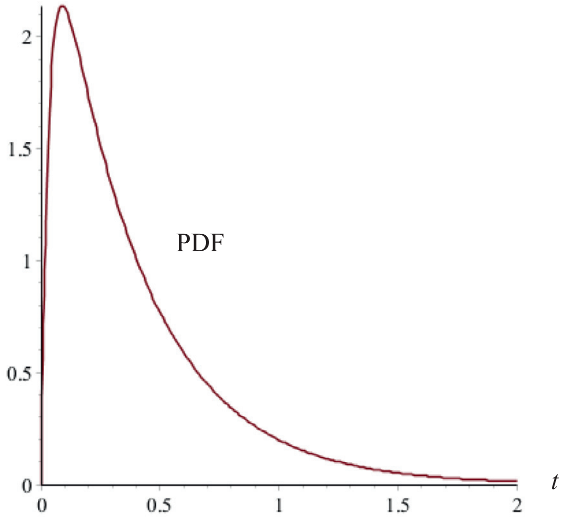

Fig. 3. Plot of $f_W(t)$ for $n = 0$, $\lambda = 8$, $\alpha = 20$, $\gamma = 18$ and $\beta = 22.5$.

7.2. The case of n = 1

Using Eqs. (18) and (19), we obtain the following performance measures:

$$L(1) = \frac{\lambda(\gamma(\gamma-\lambda)(\lambda+\alpha)+\lambda(\alpha\lambda+\gamma\beta))}{(\beta\gamma-\lambda(\beta+\gamma))(\gamma\lambda+\alpha(\gamma-\lambda))}, \tag{56}$$

and

$$L_q(1) = \frac{\lambda^2(\beta\lambda(\beta+\gamma+\alpha)+\gamma^2(\alpha+\lambda))}{\beta(\beta\gamma-\lambda(\beta+\gamma))(\gamma\lambda+\alpha(\gamma-\lambda))}. \tag{57}$$

Notice that when α is infinitely small, the server cannot produce any PS when it is idle, so the system reduces to an M/PH/1 queue, as described in the case of $n = 0$.

Using Eqs. (22) and (23), we get

$$S(1) = \frac{(\lambda+\beta)(\beta\gamma-\lambda(\beta+\gamma))\alpha}{\beta^2(\gamma\lambda+\alpha(\gamma-\lambda))}, \tag{58}$$

and

$$S_q(1) = \frac{\alpha(\beta\gamma-\lambda(\beta+\gamma))}{\beta(\gamma\lambda+\alpha(\gamma-\lambda))}. \tag{59}$$

By Little's law, $W(1)$ and $W_q(1)$ are calculated by using Eqs. (20) and (21). In order to calculate $T(1)$ and $T_q(1)$, as presented in Eqs. (24)–(26), we have to calculate the effective production rate of PSs:

$$\alpha_{eff}(1) = \alpha p_{0,0} = \frac{\alpha\lambda(\beta\gamma-\lambda(\beta+\gamma))}{\beta(\lambda\gamma+\alpha(\gamma-\lambda))}. \tag{60}$$

Substituting Eqs. (58)–(60) in Eqs. (25) and (26) results in

$$T_q(1) = 1/\lambda, \tag{61}$$

and

$$T(1) = 1/\lambda + 1/\beta. \tag{62}$$

The interpretation of Eqs. (61) and (62) is: The mean duration of time that one PS resides in inventory, $T_q(1)$, equals the mean inter-arrival time, whereas the mean duration of time that one PS resides in the system, $T(1)$, is the sum of $T_q(1)$ and the mean CS production time, $1/\beta$.

By Eq. (51), the LST of the customer's sojourn time is

$$\tilde{W}(s) = \frac{(\beta\gamma-\lambda(\beta+\gamma))(\lambda(\gamma-\alpha)+\alpha(\gamma+s))}{((\gamma+s)(\beta+s)-\lambda(\gamma+\beta+s))(\alpha\gamma+\lambda(\gamma-\alpha))}. \tag{63}$$

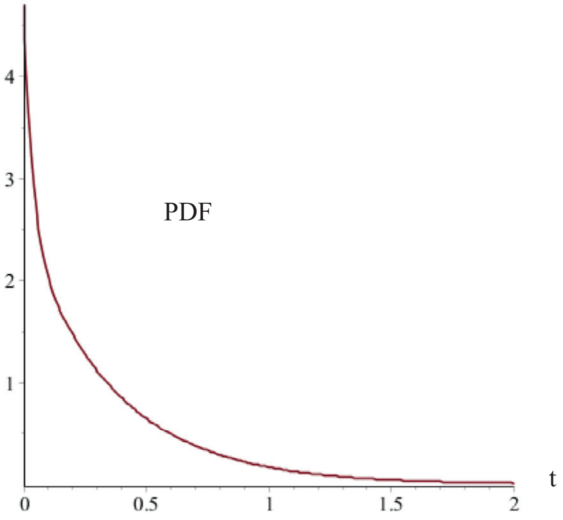

Fig. 4. Plot of $f_W(t)$ for $n = 1$, $\lambda = 8$, $\alpha = 20$, $\gamma = 18$ and $\beta = 22.5$.

Using inverse Laplace transform, we explicitly obtain the PDF of a customer's sojourn time, W:

$$f_W(t) = L^{-1}\left(\tilde{W}(s)\right) = \frac{\beta\gamma - \lambda(\gamma+\beta)}{\lambda\alpha - \gamma(\alpha+\lambda)} e^{-0.5(\gamma+\beta-\lambda)\cdot t}\left(\frac{(\alpha(\lambda+\beta-\gamma)-2\lambda\gamma)}{\Psi}\sinh(0.5\Psi\cdot t) - \alpha\cosh(0.5\Psi\cdot t)\right), \tag{64}$$

where $\cosh(x) = 0.5(e^x + e^{-x})$. In Fig. 4, we plot $f_W(t)$ for $n = 1$, $\lambda = 8$, $\alpha = 20$, $\gamma = 18$ and $\beta = 22.5$. It is observed that $f_W(t)$ is a monotone decreasing function with a right tail.

7.3. *The case of* n = 2

For simplicity of presentation, let $\mathrm{K} \equiv \beta(\lambda^3(\gamma-\alpha) - \lambda^2(\alpha-\gamma)(2\alpha+\beta) - \alpha\lambda((\beta-\gamma)\alpha - \beta\gamma) + \alpha^2\beta\gamma)$. Then

$$L(2) = \frac{\left(\begin{array}{l}(\beta(\alpha-\gamma)\lambda^4 + ((3\beta+\gamma)\alpha^2 + \beta(\beta-2\gamma)\alpha + \beta\gamma^2)\lambda^3 + ((\beta-2\gamma)\alpha^2\\ +2\alpha\gamma^2 + \beta\gamma(\beta+\gamma))\beta\lambda^2 + ((-2\beta^2\gamma+\beta\gamma^2)\alpha^2 + \beta^2\gamma^2\alpha)\lambda + \alpha^2\beta^2\gamma^2)\lambda\end{array}\right)}{\mathrm{K}(\beta\gamma-\lambda(\gamma+\beta))}, \tag{65}$$

$$S(2) = \frac{\alpha(\lambda^3 + (2\beta+3\alpha)\lambda^2 + \beta\lambda(\beta+4\alpha) + 2\alpha\beta^2)(\beta\gamma-\lambda(\gamma+\beta))}{\mathrm{K}\beta}, \tag{66}$$

$$\tilde{W}(s) = \frac{\beta\left(\begin{array}{c}(\gamma-\alpha)(\beta+s)\lambda^3 + \left(\left(\beta\gamma-\alpha^2+\alpha(2\gamma-\beta+s)\right)(\beta+s) - \alpha^2\beta\right)\lambda^2\\ -\alpha^2\beta^2\lambda + \alpha(\alpha\beta+\lambda(\alpha+\beta))(\beta+s)(\gamma+s)\end{array}\right)(\beta\gamma-\lambda(\gamma+\beta))}{\mathrm{K}(\beta+s)((\beta+s)(\gamma-\lambda+s)-\lambda\gamma)}, \tag{67}$$

and

$$\begin{aligned} f_W(t) = L^{-1}\left(\tilde{W}(s)\right) = \frac{\beta(\beta\gamma-\lambda(\gamma+\beta))}{\gamma\,\mathrm{K}\Psi} \\ \times\left[\begin{array}{l} e^{-0.5(\gamma+\beta-\lambda)\cdot t}\left(\begin{array}{r}\left(\begin{array}{l}\alpha^2\left((\beta-\gamma)\lambda^2 + \left(2\beta^2-\beta\gamma+\gamma^2\right)\lambda + \beta(\beta-\gamma)^2\right)\\ -\lambda\gamma\alpha\left(\lambda^2 + (2\beta-3\gamma)\lambda + \beta(\beta-\gamma)\right) + 2\gamma^2\lambda^2(\lambda+\beta)\end{array}\right)\sinh(0.5\Psi\cdot t)\\ +\Psi\alpha(\lambda+\beta)(\gamma(\lambda+\alpha)-\alpha\beta))\cosh(0.5\Psi\cdot t)\end{array}\right)\\ +e^{-\beta t}\alpha^2\beta(\lambda+\beta)\Psi\end{array}\right]. \end{aligned} \tag{68}$$

In Fig. 5, we plot $f_W(t)$ for $n = 2$, $\lambda = 8$, $\alpha = 20$, $\gamma = 18$ and $\beta = 22.5$. We again observe that $f_W(t)$ is a monotone decreasing function with a right tail.

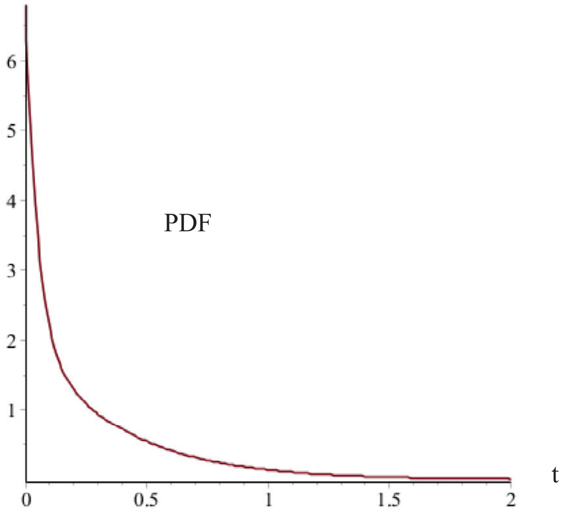

Fig. 5. Plot of $f_W(t)$ for $n = 2$, $\lambda = 8$, $\alpha = 20$, $\gamma = 18$ and $\beta = 22.5$.

8. Managerial implications

In this section, we provide managerial implications based on a practical example from the fast food industry. Consider a single-server pizzeria in which customers arrive according to a Poisson process with rate $\lambda = 5$ per hour. Stage 1 of the service is preparing the pie, which takes on average 4 min ($\gamma = 15$) when a customer is present, and 4.5 min ($\alpha = 13\frac{1}{3}$) when the pizzeria is empty. Stage 2 involves listening to the customer's requirements, composing specific ingredients (e.g., spices and extras) and putting the pizza in the oven. This stage takes on average another 4 min ($\beta = 15$). The pizza is baked for exactly $\tau = 7$ min in the oven; during this time, the server continues his work or stays idle. The pizzeria owner advertises that every customer who waits longer than $\omega = 30$ min from the time of his arrival until his pizza is ready gets a discount of 30%. Many restaurants adopt this type of compensation policy to avoid loss of goodwill; Domino's Pizza, for example, uses a 30 min guarantee policy in some countries (http://www.dominos.co.in/hot-pizza-30-minutes-delivery-guarantee-at-dominos-get-pizza-hot). A pizza is sold for $\xi = \$15$, and thus, a late pizza-supply incurs a discount of $\kappa = \$4.5$. In addition, the cost that the pizzeria owner incurs to produce a pizza unit is $c = \$5$. Since pizza is a perishable product, holding a pie in storage leads to flavor and freshness deterioration, which is associated with loss of reputation resulting in an estimated cost of $h = \$0.25$ per storage hour per pizza.

The owner's objective is to maximize its expected profit per hour by controlling the maximal number of stored pizza pies. The monetary objective function comprises three elements: (i) revenue minus production cost; (ii) loss due to flavor deterioration, which is proportional (with coefficient h) to the expected number of stored pizza pies; and (iii) loss due to late pizza-supply discounts. Accordingly, the owner's problem is formulated as:

$$\max_n \left\{ Z(n) = \lambda(\xi - c) - h \cdot S_q(n) - \lambda \cdot \kappa \int_{\omega-\tau}^{\infty} f_W(t)dt \right\}. \tag{69}$$

In Eq. (69), there is a tradeoff between the two cost components as a function of n. When n increases, the inventory level of pizza pies, $S_q(n)$, increases, resulting in higher loss due to flavor deterioration, whereas the volume of sales at a discount, $\lambda \int_{\omega-\tau}^{\infty} f_W(t)dt$, decreases. In the example above, the optimal solution is $n^* = 7$, which results in $Z(7) = \$47.18$. A policy in which the server uses his idle time to produce and store up to seven pizza pies leads to an increase of 9.85% in the owner's expected profit compared with a policy in which the server's idle time is not utilized at all ($Z(0) = \$42.94$).

As a sensitivity analysis, Fig. 6 presents the expected profit of the pizzeria owner as a function of the maximal number of stored pizza pies, n, for three values of $h \in \{\$0.1, \$0.25, \$0.4\}$. As anticipated, the expected profit is a concave function of n, and higher values of h result in lower optimal values of n ($n^* = \{10, 7, 5\}$, marked with circles on the curves). Indeed, when customers are more sensitive to flavor deterioration, keeping pizza pies in inventory is less beneficial.

We now consider a more general case, in which the value of the discount for a late pizza-supply affects the average arrival rate of customers. This model introduces a new kind of tradeoff: Increasing the discount for a late pizza-supply increases the expected gross profit, $\lambda(\kappa)\cdot(\xi - c)$, due to an increase in customers' arrival rate. However, it also increases the loss due to lateness, since the server has less idle time to produce pizza pies, so the probability of a late pizza-supply increases, while at the same time customers' arrival rate is higher and the discount value is greater. We assume that λ is an increasing concave function of κ, according to the law of diminishing marginal returns. Thus, the owner's problem is

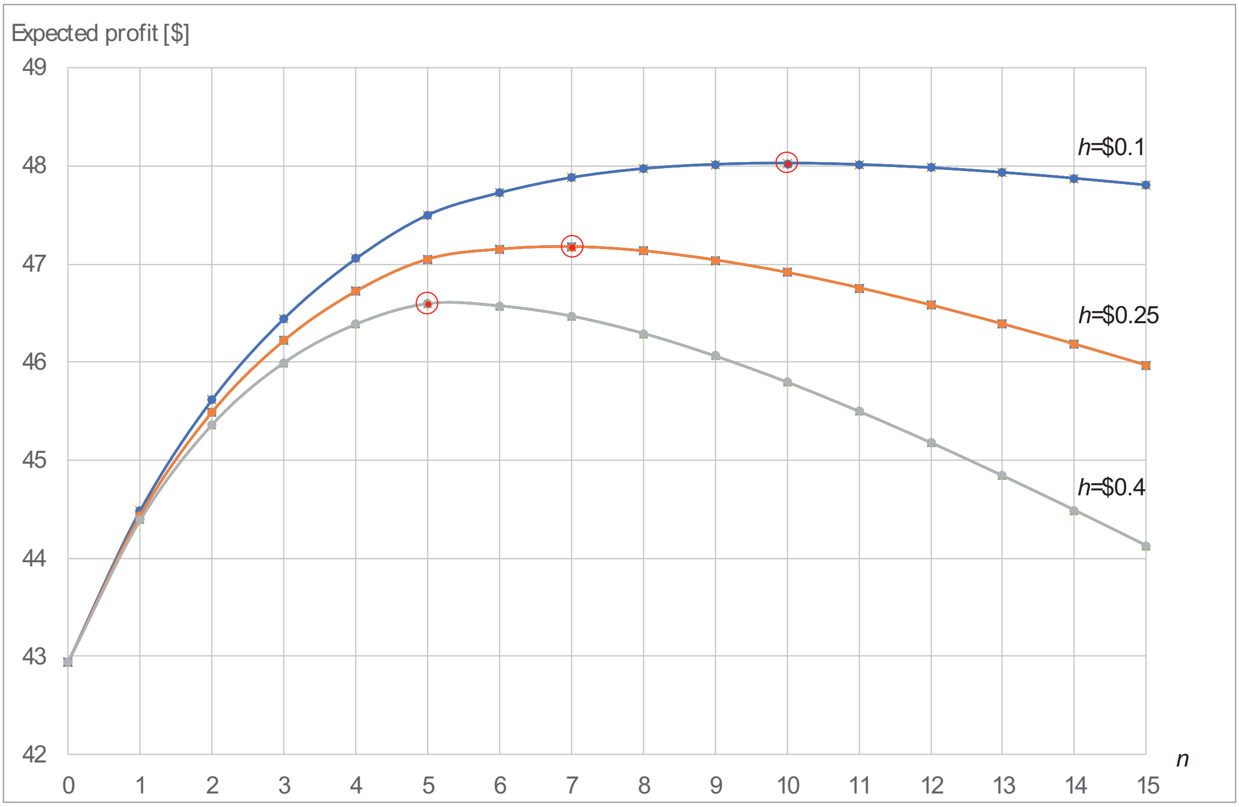

Fig. 6. Expected profit as a function of n for $h = \{\$0.1, \$0.25, \$0.4\}$, where $\alpha = 13\frac{1}{3}$, $\beta = 15$, $\gamma = 15$, $\lambda = 15$, $\omega = 30$, $\tau = 7$, $\xi = \$15$, $\kappa = \$4.5$ and $c = \$5$.

Table 1
Expected profit (in \$) as a function of n and selected values of κ for $\lambda(\kappa) = 5 - \exp(-\kappa)$, where $\alpha = 13\frac{1}{3}, \beta = 15$, $\gamma = 15$, $\lambda = 15$, $\omega = 30$, $\tau = 7$, $\xi = \$15$, $h = \$0.25$ and $c = \$5$.

$n\backslash\kappa$	0	0.5	1	1.5	2	2.5	**3**	3.5	4	4.5	5	5.5	6	6.5	7
0	40.00	43.42	45.10	45.75	45.79	45.47	44.95	44.32	43.62	42.94	42.13	41.36	40.58	39.80	39.02
1	39.89	43.47	45.33	46.16	46.38	46.25	45.91	45.45	44.93	44.44	43.79	43.20	42.60	42.00	41.39
2	39.71	43.41	45.41	46.38	46.74	46.75	46.55	46.23	45.85	45.49	44.97	44.51	44.04	43.57	43.10
3	39.51	43.29	45.39	46.47	46.94	47.06	46.96	46.75	46.47	46.22	45.80	45.44	45.07	44.70	44.32
4	39.29	43.13	45.31	46.47	47.03	47.23	47.22	47.09	46.89	46.72	46.37	46.09	45.80	45.50	45.20
5	39.07	42.95	45.19	46.41	47.04	47.31	**47.36**	47.29	47.15	47.05	46.76	46.53	46.30	46.06	45.82
6	38.83	42.74	45.01	46.27	46.94	47.25	47.34	47.31	47.21	47.15	46.89	46.70	46.50	46.30	46.10
7	38.59	42.51	44.81	46.11	46.81	47.15	47.28	47.28	47.20	47.18	46.94	46.78	46.61	46.43	46.25
8	38.34	42.28	44.59	45.92	46.65	47.01	47.16	47.19	47.13	47.13	46.91	46.77	46.62	46.47	46.31
9	38.10	42.04	44.37	45.71	46.46	46.85	47.02	47.06	47.02	47.04	46.83	46.71	46.58	46.44	46.29
10	37.85	41.80	44.14	45.50	46.26	46.66	46.85	46.90	46.88	46.91	46.72	46.60	46.48	46.36	46.22
11	37.60	41.56	43.91	45.27	46.05	46.46	46.66	46.72	46.71	46.76	46.57	46.47	46.35	46.23	46.11
12	37.36	41.31	43.67	45.04	45.83	46.25	46.45	46.53	46.53	46.58	46.40	46.30	46.20	46.09	45.97
13	37.11	41.06	43.42	44.81	45.60	46.03	46.24	46.32	46.32	46.39	46.21	46.12	46.02	45.91	45.80
14	36.86	40.82	43.18	44.57	45.36	45.80	46.02	46.11	46.11	46.18	46.01	45.92	45.83	45.72	45.62
15	36.61	40.57	42.93	44.33	45.13	45.57	45.79	45.88	45.89	45.97	45.79	45.71	45.62	45.52	45.42

formulated as:

$$\max_{n,\kappa}\left\{Z(n,\kappa) = \lambda(\kappa)\cdot(\xi - c) - h\cdot S_q(n) - \lambda(\kappa)\cdot\kappa\int_{\omega-\tau}^{\infty} f_W(t)dt\right\}. \tag{70}$$

For illustration purposes, we use $\lambda(\kappa) = 5 - \exp(-\kappa)$, so that $\lambda(4.5) \approx 5$ (to be consistent with the previous example). Fig. 7 presents the expected profit of the pizzeria owner as a function of the maximal number of stored pizza pies, n, and the

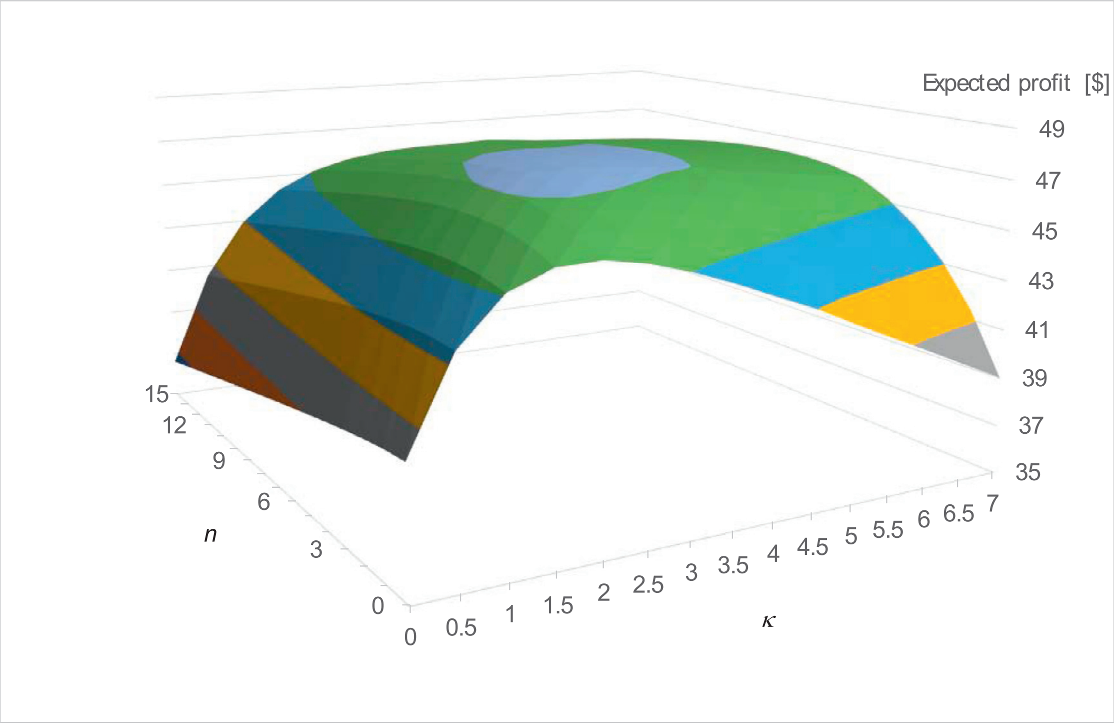

Fig. 7. Expected profit as a function of n and κ for $\lambda(\kappa) = 5 - \exp(-\kappa)$, where $\alpha = 13\frac{1}{3}$, $\beta = 15$, $\gamma = 15$, $\lambda = 15$, $\omega = 30$, $\tau = 7$, $\xi = \$15$, $h = \$0.25$ and $c = \$5$.

discount given for a late pizza-supply, κ. The numerical values are given in Table 1 for discrete values of κ. As anticipated, the expected profit is jointly concave in n and κ, and the maximal expected profit is obtained at $n = 5$ and $\kappa \approx 3$.

In the example above, the optimal solution is $n^* = 5$ and $\kappa^* = 3$, which results in $Z(5, 3) = \$47.36$. Using the server's idle time to produce and store up to five pizza pies and offering a discount of \$3 per late pizza-supply leads to an increase of 18.4% in the owner's expected profit compared with a policy in which the server's idle time is not utilized and in which no discount is offered for a late pizza-supply ($Z(0, 0) = \$40$). These outcomes suggest that our model is applicable and can help decision makers in achieving higher profits.

9. Conclusions and discussion

In this paper, we have considered a single-server queue in which customers arrive according to a Poisson process, and the service provided to each individual customer is composed of two independent stages. The first stage can be carried out either ahead of the customer's arrival or in his/her presence, whereas the second stage requires the customer to be present. The service duration of each stage is distributed exponentially, but the duration of the first stage may depend on whether the customer is present or absent while the service is being performed. Our novel approach entails letting the server utilize its idle time to produce PSs that can be stored in inventory, with the goal of improving the system's performance. We formulated the queueing-inventory system as a QBD process, and explicitly derived the entries of the rate matrix R. Consequently, we derived explicit expressions for various performance measures. With the use of Maple 2016 software, we analytically solved problems with large values of the inventory capacity level in a relatively short computational time. We also showed that the stability condition of the system is independent of the PS production rate.

We found that the mean sojourn time of a customer is a monotone decreasing convex function of the PS capacity, and approaches an asymptotic value. This observation indicates that decomposing a service into two stages and storing PSs is beneficial for the system's efficiency. In some cases, the mean values of performance measures do not fully reflect the system's performance, and exact distributions of system variables are required. Hence, we derived explicitly the distribution of a customer's sojourn time by using the Laplace–Stieltjes transform and its inverse.

We carried out an economic analysis using a practical example from the fast food industry, and found the optimal inventory capacity of PSs as well as the optimal pricing policy. Future work might extend our model in the following directions: (i) studying non-Markovian queues; (ii) investigating multi-server queues; and (iii) extending the two-stage service to a multi-stage service. Those directions are likely to require simulation analysis due to their analytical complexity.

Appendix A

$$A_0+RA_1+R^2A_2=$$

$$\begin{pmatrix}
\lambda-(\lambda+\beta)r_{1,1}+\gamma r_{1,2} & -(\lambda+\gamma)r_{1,2}+\beta\sum\limits_{k=1}^{n+2}r_{1,k}(r_{k,1}+r_{k,3}) & -(\lambda+\beta)r_{1,3}+\beta\sum\limits_{k=1}^{n+2}r_{1,k}r_{k,4} & -(\lambda+\beta)r_{1,4}+\beta\sum\limits_{k=1}^{n+2}r_{1,k}r_{k,5} & \cdots & -(\lambda+\beta)r_{1,n}+\beta\sum\limits_{k=1}^{n+2}r_{1,k}r_{k,n+1} & -(\lambda+\beta)r_{1,n+1}+\beta\sum\limits_{k=1}^{n+2}r_{1,k}r_{k,n+2} & -(\lambda+\beta)r_{1,n+2} \\
-(\lambda+\beta)r_{2,1}+\gamma r_{2,2} & \lambda-(\lambda+\gamma)r_{2,2}+\beta\sum\limits_{k=1}^{n+2}r_{2,k}(r_{k,1}+r_{k,3}) & -(\lambda+\beta)r_{2,3}+\beta\sum\limits_{k=1}^{n+2}r_{2,k}r_{k,4} & -(\lambda+\beta)r_{2,4}+\beta\sum\limits_{k=1}^{n+2}r_{2,k}r_{k,5} & \cdots & -(\lambda+\beta)r_{2,n}+\beta\sum\limits_{k=1}^{n+2}r_{2,k}r_{k,n+1} & -(\lambda+\beta)r_{2,n+1}+\beta\sum\limits_{k=1}^{n+2}r_{2,k}r_{k,n+2} & -(\lambda+\beta)r_{2,n+2} \\
-(\lambda+\beta)r_{3,1}+\gamma r_{3,2} & -(\lambda+\gamma)r_{3,2}+\beta\sum\limits_{k=1}^{n+2}r_{3,k}(r_{k,1}+r_{k,3}) & \lambda-(\lambda+\beta)r_{3,3}+\beta\sum\limits_{k=1}^{n+2}r_{3,k}r_{k,4} & -(\lambda+\beta)r_{3,4}+\beta\sum\limits_{k=1}^{n+2}r_{3,k}r_{k,5} & \cdots & -(\lambda+\beta)r_{3,n}+\beta\sum\limits_{k=1}^{n+2}r_{3,k}r_{k,n+1} & -(\lambda+\beta)r_{3,n+1}+\beta\sum\limits_{k=1}^{n+2}r_{3,k}r_{k,n+2} & -(\lambda+\beta)r_{3,n+2} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
-(\lambda+\beta)r_{n,1}+\gamma r_{n,2} & -(\lambda+\gamma)r_{n,2}+\beta\sum\limits_{k=1}^{n+2}r_{n,k}(r_{k,1}+r_{k,3}) & -(\lambda+\beta)r_{n,3}+\beta\sum\limits_{k=1}^{n+2}r_{n,k}r_{k,4} & -(\lambda+\beta)r_{n,4}+\beta\sum\limits_{k=1}^{n+2}r_{n,k}r_{k,5} & \cdots & \lambda-(\lambda+\beta)r_{n,n}+\beta\sum\limits_{k=1}^{n+2}r_{n,k}r_{k,n+1} & -(\lambda+\beta)r_{n,n+1}+\beta\sum\limits_{k=1}^{n+2}r_{n,k}r_{k,n+2} & -(\lambda+\beta)r_{n,n+2} \\
-(\lambda+\beta)r_{n+1,1}+\gamma r_{n+1,2} & -(\lambda+\gamma)r_{n+1,2}+\beta\sum\limits_{k=1}^{n+2}r_{n+1,k}(r_{k,1}+r_{k,3}) & -(\lambda+\beta)r_{n+1,3}+\beta\sum\limits_{k=1}^{n+2}r_{n+1,k}r_{k,4} & -(\lambda+\beta)r_{n+1,4}+\beta\sum\limits_{k=1}^{n+2}r_{n+1,k}r_{k,5} & \cdots & -(\lambda+\beta)r_{n+1,n}+\beta\sum\limits_{k=1}^{n+2}r_{n+1,k}r_{k,n+1} & \lambda-(\lambda+\beta)r_{n+1,n+1}+\beta\sum\limits_{k=1}^{n+2}r_{n+1,k}r_{k,n+2} & -(\lambda+\beta)r_{n+1,n+2} \\
-(\lambda+\beta)r_{n+2,1}+\gamma r_{n+2,2} & -(\lambda+\gamma)r_{n+2,2}+\beta\sum\limits_{k=1}^{n+2}r_{n+2,k}(r_{k,1}+r_{k,3}) & -(\lambda+\beta)r_{n+2,3}+\beta\sum\limits_{k=1}^{n+2}r_{n+2,k}r_{k,4} & -(\lambda+\beta)r_{n+2,4}+\beta\sum\limits_{k=1}^{n+2}r_{n+2,k}r_{k,5} & \cdots & -(\lambda+\beta)r_{n+2,n}+\beta\sum\limits_{k=1}^{n+2}r_{n+2,k}r_{k,n+1} & -(\lambda+\beta)r_{n+2,n+1}+\beta\sum\limits_{k=1}^{n+2}r_{n+2,k}r_{k,n+2} & \lambda-(\lambda+\beta)r_{n+2,n+2}
\end{pmatrix}. \tag{1.1}$$

Appendix B

(i) We first prove by induction that $r_{i,\,j}=0,\ 3\le j\le n+2\,,\ \ i<j$. Starting with the last column of Eq. (1.1) above, i.e., column $j=n+2$, we obtain from Eq. (11):

$$-(\lambda+\beta)r_{i,n+2}=0,\ \ i=1,2,...,n+1, \tag{2.1}$$

implying that $r_{i,\,n+2}=0$ for $i=1,2,...,n+1$.
We now assume that $r_{i,\,j}=0$ for all $1\le i<j\le n+2, j\ge n+2-m, 0\le m\le n-1$, and show that $r_{i,\,n+1-m}=0$ for $1\le i\le n-m$. By Eq. (11) and the entries above the main diagonal in column $j=n+1-m$ of Eq. (1.1), $-(\lambda+\beta)r_{i,n+1-m}+\beta\sum_{k=1}^{n+2}r_{i,k}r_{k,n+2-m}=0,\ i=1,2,...,n-m$. By the induction assumption, $r_{k,\,n+2-m}=0$ for $1\le k\le n+1-m$, and $r_{i,\,k}=0$ for $n+1-m<k\le n+2$ since $i\le n-m$. Thus, $\sum_{k=1}^{n+2}r_{i,k}r_{k,n+2-m}=0$. Hence, $r_{i,\,n+1-m}=0$ for $1\le i\le n-m$, which proves the claim.

(ii) We now prove by induction that $r_{i,j}=\frac{C_{i-j}\beta^{i-j}\lambda^{i-j+1}}{(\lambda+\beta)^{2(i-j)+1}}$, $3\le j\le i\le n+2$, where C_m is the mth Catalan number. Starting with the main diagonal of (1.1), i.e., $i=j\ge 3$, we have from Eq. (11):

$$\lambda-(\lambda+\beta)r_{i,i}+\beta\sum_{k=1}^{n+2}r_{i,k}r_{k,i+1}=0,\ \ i=3,4,...,n+1, \tag{2.2}$$

$$\lambda-(\lambda+\beta)r_{n+2,n+2}=0. \tag{2.3}$$

By (i), $r_{k,\,i+1}=0$ for $1\le k\le i$, and $r_{i,\,k}=0$ for $i+1\le k\le n+2$, so $\sum_{k=1}^{n+2}r_{i,k}r_{k,i+1}=0$. Thus, from Eqs. (2.2) and (2.3),

$$r_{i,i}=\lambda/(\lambda+\beta)=\frac{C_0\beta^0\lambda^1}{(\lambda+\beta)^{2(i-i)+1}},\ \ i=3,4,...,n+2.$$

Next, we assume that $r_{i,j}=\frac{C_{i-j}\beta^{i-j}\lambda^{i-j+1}}{(\lambda+\beta)^{2(i-j)+1}}$ for all $0\le i-j\le m\le n-2,\ 3\le j\le i\le n+2$ (i.e., expressions for the entries in the main diagonal and in the $m-1$ diagonals below it) and show that $r_{i,j}=\frac{C_{i-j}\beta^{i-j}\lambda^{i-j+1}}{(\lambda+\beta)^{2(i-j)+1}}$ for $i-j=m+1$,

$3 \le j \le i \le n+2$. By Eq. (11) and the entries in the diagonal $i-j=m+1$ of Eq. (1.1),

$$-(\lambda+\beta)r_{i,i-m-1}+\beta\sum_{k=1}^{n+2} r_{i,k}r_{k,i-m}=0,\ i=m+3,\ldots,n+2. \tag{2.4}$$

From (i), $r_{k,\,i-m}=0$ for $1 \le k \le i-m-1$ and $r_{i,\,k}=0$ for $i+1 \le k \le n+2$, so Eq. (2.4) can be written as $-(\lambda+\beta)r_{i,i-m-1}+\beta\sum_{k=i-m}^{i} r_{i,k}r_{k,i-m}=0$, $i=m+3, \ldots, n+2$. By the induction assumption, $r_{i,k}=\frac{C_{i-k}\beta^{i-k}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)+1}}$ for $i-k \le m$ (and thus for $i-m \le k \le i$) and $r_{k,i-m}=\frac{C_{k-i+m}\beta^{k-i+m}\lambda^{k-i+m+1}}{(\lambda+\beta)^{2(k-i+m)+1}}$ for $k-i+m \le m$ (and thus for $i-m \le k \le i$), so $\sum_{k=i-m}^{i} r_{i,k}r_{k,i-m}=\frac{\beta^{m}\lambda^{m+2}}{(\lambda+\beta)^{2(m+1)}}\sum_{k=i-m}^{i} C_{i-k}C_{k-i+m}$. By modifying the indices in the summation, $\sum_{k=i-m}^{i} C_{i-k}C_{k-i+m}=\sum_{k=0}^{m} C_kC_{,m-k}$, and from using the recurrence relation of the Catalan numbers, $\sum_{k=0}^{m} C_kC_{m-k}=C_{m+1}$. Thus, Eq. (2.4) can be written as

$$-(\lambda+\beta)r_{i,i-m-1}+\frac{C_{m+1}\beta^{m+1}\lambda^{m+2}}{(\lambda+\beta)^{2(m+1)}}=0,\ i=m+3,\ldots,n+2, \tag{2.5}$$

from which the claim is proved, i.e., $r_{i,j}=\frac{C_{i-j}\beta^{i-j}\lambda^{i-j+1}}{(\lambda+\beta)^{2(i-j)+1}}$, $i-j=m+1$, $3 \le j \le i \le n+2$.

(iii) Next we show that $r_{1,\,1}=r_{2,\,1}=\lambda/\beta$, $r_{1,\,2}=\lambda^2/\gamma\beta$ and $r_{2,\,2}=\lambda(\lambda+\beta)/\gamma\beta$. By the two first entries in the two first columns of Eq. (1.1) above, we obtain the following equations:

$$\lambda-(\lambda+\beta)r_{1,1}+\gamma\, r_{1,2}=0, \tag{2.6}$$

$$-(\lambda+\beta)r_{2,1}+\gamma\, r_{2,2}=0, \tag{2.7}$$

$$-(\lambda+\gamma)r_{1,2}+\beta\sum_{k=1}^{n+2} r_{1,k}(r_{k,1}+r_{k,3})=0, \tag{2.8}$$

$$\lambda-(\lambda+\gamma)r_{2,2}+\beta\sum_{k=1}^{n+2} r_{2,k}(r_{k,1}+r_{k,3})=0. \tag{2.9}$$

By (i), $r_{1,\,k}=0$ for $3 \le k \le n+2$ and $r_{k,\,3}=0$ for $1 \le k \le 2$. Thus, Eqs. (2.8) and (2.9) are reduced to

$$-(\lambda+\gamma)r_{1,2}+\beta r_{1,1}^2+\beta r_{1,2}r_{2,1}=0, \tag{2.10}$$

$$\lambda-(\lambda+\gamma)r_{2,2}+\beta r_{2,1}r_{1,1}+\beta r_{2,2}r_{2,1}=0. \tag{2.11}$$

Solving Eqs. (2.6), (2.7), (2.10) and (2.11) results in $r_{1,\,1}=r_{2,\,1}=\lambda/\beta$, $r_{1,\,2}=\lambda^2/\gamma\beta$ and $r_{2,\,2}=\lambda(\lambda+\beta)/\gamma\beta$.

(iv) Now we show that for $i=3, 4, \ldots, n+2$, $r_{i,2}=\frac{C_{i-2}\beta^{i-4}\lambda^{i-1}}{\gamma(\lambda+\beta)^{2i-6}}+\sum_{k=3}^{i-1}\frac{C_{i-k}\beta^{i-k-1}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)}}r_{k,1}$ and $r_{i,1}=\frac{C_{i-2}\beta^{i-4}\lambda^{i-1}}{(\lambda+\beta)^{2i-5}}+\sum_{k=3}^{i-1}\frac{C_{i-k}\beta^{i-k-1}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)}}r_{k,1}$. By the entries in the first two columns of Eq. (1.1) above, except for those in the first two rows, we obtain the following equations:

$$-(\lambda+\gamma)r_{i,2}+\beta\sum_{k=1}^{n+2} r_{i,k}(r_{k,1}+r_{k,3})=0,\ i=3,4,\ldots,n+2, \tag{2.12}$$

$$-(\lambda+\beta)r_{i,1}+\gamma r_{i,2}=0,\ i=3,4,\ldots,n+2. \tag{2.13}$$

By (i) $r_{i,\,k}=0$ for $i+1 \le k \le n+2$ and $r_{k,\,3}=0$ for $1 \le k \le 2$. Thus, Eq. (2.12) can be written as

$$-(\lambda+\gamma)r_{i,2}+\beta\left(r_{i,1}r_{1,1}+r_{i,2}r_{2,1}+\sum_{k=3}^{i-1} r_{i,k}r_{k,1}+r_{i,i}r_{i,1}\right)+\beta\sum_{k=3}^{i} r_{i,k}r_{k,3}=0,\ i=3,4,\ldots,n+2. \tag{2.14}$$

By (ii) $r_{i,k}=\frac{C_{i-k}\beta^{i-k}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)+1}}$ and $r_{k,3}=\frac{C_{k-3}\beta^{k-3}\lambda^{k-2}}{(\lambda+\beta)^{2(k-3)+1}}$ for $3 \le k \le i$; by (iii) $r_{1,\,1}=\lambda/\beta$ and $r_{2,\,1}=\lambda/\beta$; by Eq. (2.13) $r_{i,\,1}=\gamma r_{i,\,2}/(\lambda+\beta)$ for $3 \le i \le n+2$ and $r_{k,\,1}=\gamma r_{k,\,2}/(\lambda+\beta)$ for $3 \le k \le i$. Thus, Eq. (2.14) can be written as

$$\frac{-\gamma\beta^2}{(\lambda+\beta)^2} r_{i,2} + \frac{\beta^{i-2}\lambda^{i-1}}{(\lambda+\beta)^{2(i-2)}} \sum_{k=3}^{i} C_{i-k}C_{k-3} + \beta \sum_{k=3}^{i-1} \frac{C_{i-k}\gamma\beta^{i-k}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)+2}} r_{k,2} = 0,\ \ i = 3, 4, ..., n+2. \tag{2.15}$$

Multiplying Eq. (2.15) by $(\lambda+\beta)/(\gamma\beta^2)$ and replacing $\sum_{k=3}^{i} C_{i-k}C_{k-3}$ with $\sum_{k=0}^{i-3} C_{i-3-k}C_k = C_{i-2}$ results in

$$r_{i,2} = \frac{C_{i-2}\beta^{i-4}\lambda^{i-1}}{\gamma(\lambda+\beta)^{2(i-3)}} + \sum_{k=3}^{i-1} \frac{C_{i-k}\beta^{i-k-1}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)}} r_{k,2},\ \ i = 3, 4, ..., n+2. \tag{2.16}$$

Substituting Eq. (2.16) in (2.13) results in

$$r_{i,1} = \frac{C_{i-2}\beta^{i-4}\lambda^{i-1}}{(\lambda+\beta)^{2i-5}} + \sum_{k=3}^{i-1} \frac{C_{i-k}\beta^{i-k-1}\lambda^{i-k+1}}{(\lambda+\beta)^{2(i-k)}} r_{k,1},\ \ i = 3, 4, ..., n+2. \tag{2.17}$$

This completes the proof.

References

[1] Y. Levy, U. Yechiali, Utilization of idle time in an M/G/1 queueing system, Manag. Sci. 22 (1975) 202–211.

[2] Y. Levy, U. Yechiali, An M/M/s queue with servers' vacations, INFOR 14 (1976) 153–163.

[3] O. Kella, U. Yechiali, Priorities in M/G/1 queue with server vacations, Nav. Res. Logist. 35 (1988) 23–34.

[4] H. Takagi, Queueing Analysis, Volume 1: Vacation and Priority Systems, North-Holland, Amsterdam, 1991.

[5] E. Rosenberg, U. Yechiali, The $M^X/G/1$ queue with single and multiple vacations under the LIFO service regime, Oper. Res. Lett. 14 (3) (1993) 171–179.

[6] O.J. Boxma, S. Schlegel, U. Yechiali, A note on the M/G/1 queue with a waiting server, timer and vacations, Am. Math. Soc. Transl. Ser. 2 (2002) 25–35.

[7] U. Yechiali, On the $M^X/G/1$ queue with a waiting server and vacations, Sankhya 66 (2004) 159–174.

[8] C.H. Lin, J.C. Ke, Multi-server system with single working vacation, Appl. Math. Model. 33 (7) (2009) 2967–2977.

[9] M. Jain, A. Jain, Working vacations queueing model with multiple types of server breakdowns, Appl. Math. Model. 34 (1) (2010) 1–13.

[10] J.C. Ke, C.H. Wu, W.L. Pearn, Algorithmic analysis of the multi-server system with a modified Bernoulli vacation schedule, Appl. Math. Model. 35 (5) (2011) 2196–2208.

[11] G.C. Mytalas, M.A. Zazanis, An MX/G/1 queueing system with disasters and repairs under a multiple adapted vacation policy, Nav. Res. Logist. 62 (2015) 171–189.

[12] D. Guha, V. Goswami, A.D. Banik, Algorithmic computation of steady-state probabilities in an almost observable GI/M/c queue with or without vacations under state dependent balking and reneging, Appl. Math. Model. 40 (5) (2016) 4199–4219.

[13] M. Yadin, P. Naor, Queueing systems with a removable service station, Oper. Res. Q. 14 (1963) 393–405.

[14] O. Kella, The threshold policy in the M/G/1 queue with server vacations, Nav. Res. Logist. 36 (1989) 111–123.

[15] P. Moreno, A discrete-time single-server queue with a modified N-policy, Int. J. of Syst. Sci. 38 (2007) 483–492.

[16] D.H. Lee, W.S. Yang, The N-policy of a discrete time Geo/G/1 queue with disasters and its application to wireless sensor networks, Appl. Math. Model. 37 (2013) 9722–9731.

[17] D.E. Lim, D.H. Lee, W.S. Yang, K.C. Chae, Analysis of the GI/Geo/1 queue with N-policy, Appl. Math. Model. 37 (2013) 4643–4652.

[18] Y. Wei, M. Yu, Y. Tang, J. Gu, Queue size distribution and capacity optimum design for N-policy Geo (λ1, λ2, λ3)/G/1 queue with setup time and variable input rate, Math. Comput. Model. 57 (2013) 1559–1571.

[19] D. Yang, C.H. Wu, Cost-minimization analysis of a working vacation queue with N-policy and server breakdowns, Comput. Ind. Eng. 82 (2015) 151–158.

[20] M. Haridass, R. Arumuganathan, Analysis of a single server batch arrival retrial queueing system with modified vacations and N-policy, RAIRO-Oper. Res. 49 (2015) 279–296.

[21] G. Choudhury, K. Madan, A two-stage batch arrival queueing system with a modified Bernoulli schedule vacation under N-policy, Math. Comput. Model. 42 (2005) 71–85.

[22] G. Choudhury, A two phase batch arrival retrial queueing system with Bernoulli vacation schedule, Appl. Math. Comput. 188 (2007) 1455–1466.

[23] G. Choudhury, L. Tadj, M. Paul, Steady state analysis of an M x/G/1 queue with two phase service and Bernoulli vacation schedule under multiple vacation policy, Appl. Math. Model. 31 (2007) 1079–1091.

[24] G. Choudhury, Steady state analysis of an M/G/1 queue with linear retrial policy and two phase service under Bernoulli vacation schedule, Appl. Math. Model. 32 (2008) 2480–2489.

[25] G. Choudhury, M. Deka, A single server queueing system with two phases of service subject to server breakdown and Bernoulli vacation, Appl. Math. Model. 36 (2012) 6050–6060.

[26] N. Litvak, U. Yechiali, Routing in Queues with Delayed Information, Queueing Syst. 43 (2003) 147–165.

[27] E. Perel, U. Yechiali, Queues where customers of one queue act as servers of the other queue, Queueing Syst. 60 (2008) 271–288.

[28] N. Perel, U. Yechiali, The Israeli queue with retrials, Queueing Syst. 78 (2014) 31–56.

[29] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, Johns Hopkins University Press, Baltimore, MD, 1981.

[30] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, ASA-SIAM Series On Statistics and Applied Probability, SIAM, Philadelphia, PA, 1999.

[31] G. Hanukov, T. Avinadav, T. Chernonog, U. Spiegel, U. Yechiali, Improving Efficiency in Queueing Systems By Utilizing the Server's Idle Time, Bar-Ilan University, 2016 Working paper.

[32] Y. Ma, W.Q. Liu, J.H. Li, Equilibrium balking behavior in the Geo/Geo/1 queueing system with multiple vacations, Appl. Math. Model. 37 (6) (2013) 3861–3878.

[33] W. Zhou, W. Huang, R. Zhang, A two-stage queueing network on form postponement supply chain with correlated demands, Appl. Math. Model. 38 (11) (2014) 2734–2743.

[34] T. Koshy, Catalan Numbers with Applications, Oxford University Press, Oxford, 2008.

[35] R.B. Cooper, Introduction to Queueing Theory, North Holland, New York, Oxford, 1981.